DOCUMENT RESUME

ED 300 453 TM 012 503

AUTHOR Kingston, Neal
TITLE Analysis of Shifts in Scale and Construct through the Use of Repeater Data.
SPONS AGENCY Graduate Record Examinations Board, Princeton, N.J.
PUB DATE Apr 84
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 23-27, 1984).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *College Entrance Examinations; Equated Scores; Higher Education; Mathematics Tests; *Multidimensional Scaling; Rating Scales; *Scoring; Scoring Formulas; *Test Construction; Testing Problems; *Timed Tests; Verbal Tests
IDENTIFIERS Analytical Tests; *Graduate Record Examinations; Response Shift; Rights and Formula Scoring; Test Repeaters; Test Revision

ABSTRACT

In October 1981, the Graduate Record Examinations (GRE) Program introduced a new version of the General Test (GT) that differed from the previous version in three major ways. The GT was altered to: reduce the verbal measure's speededness and allow the addition of several quantitative items; delete two item types from the analytical measure; and replace formula scoring with rights scoring. The GRE instructions were changed to advise examinees to answer all questions. An anchor test design was used to equate verbal and quantitative measures. Focus was on determining what effects these changes and/or the equating (in the case of the verbal and quantitative measures) or the scaling (in the case of the analytical measure) had on the three GRE GT score scales--verbal, quantitative, and analytical. These changes could be manifested as either shifts in the score scales or changes in the constructs underlying the scales. Data were from a self-selected group of GRE examinees who took the GT for the first time between October 1980 and April 1982 and at least one additional time by June 1982. A total of 5,072 GRE examinees took the old version at least twice, and 2,353 took the old version first and then the new version. The 1981 changes had a small effect on the verbal score scale and a somewhat greater effect on the quantitative scale; these changes are attributed to a shift in the dimensionality of the factors underlying these measures. The effect on the analytical scale was quite large. Five data tables and three graphs conclude the document. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED300453

Analysis of Shifts in Scale and Construct Through the Use of Repeater Data

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

NEAL M. KINGSTON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Neal Kingston
Educational Testing Service

A paper presented at the annual meeting of the American Educational Research Association, New Orleans, April 24, 1984. Support for this research was provided by the Graduate Record Examinations Board. The views expressed herein, however, are solely those of the author.

TM012 503

INTRODUCTION

Since October 1977 the Graduate Record Examinations (GRE) General Test, referred to through 1981 as the GRE Aptitude Test, has consisted of three measures of developed abilities: verbal, quantitative, and analytical. In October 1981 the Graduate Record Examinations Program introduced a new version of the General Test that differed from the previous version in three major ways. A change in timing was designed to reduce the speededness of the verbal measure and to allow the addition of several quantitative items. Although the content of the verbal and quantitative measures remained essentially the same, the analytical measure was significantly revised by deleting two item types that had shown a short term (within test-form) practice effect and a susceptibility to coaching. In addition to these changes, the scoring for all three measures was changed from formula scoring (subtracting a fraction of the number of incorrect responses from the number of questions answered correctly) to rights scoring (counting only the number of questions answered correctly). Instructions to the examinees were correspondingly changed: examinees were advised to answer every question.

Although the GRE General Test is usually equated by administering a new edition and an old edition of the test to equivalent groups and setting means and standard deviations equal (Equating Design I, Angoff, 1971), the change in test format made this administratively not feasible for the editions introduced in October 1981, so an anchor test design was used (Equating Design IV, Angoff, 1971) to equate the verbal and quanitative measures. Since the changes to the analytical measure were more major, and since the measure was and still is considered experimental, no attempt was made to equate it to the old format analytical measure scale. Instead, it was placed on scale through scores on the verbal and quantitative measures as had originally been done for the earlier analytical measure in 1977. The anchor test design used to equate the verbal and quantitative measures appears to have been the best design that was administratively feasible, but, the assumptions of the design were not met.

This paper will try to ascertain what effects, if any, the changes instituted in October 1981 and/or the equating (in the case of the verbal and quantitative measures) or the scaling (in the case of the analytical measure) had on the three GRE General Test score scales. These changes could be manifested as either shifts in the score scales or changes in the constructs underlying the scales.

RESEARCH DESIGN

Data base

The analyses are based on the self-selected group of Graduate Record Examinations test takers who took the General Test the first time between October 1980 and April 1982 and took the test at least one additional time by June 1982. Only those examinees taking the test in the pairs of domestic administrations indicated in Table 1 were included. Only the scores, verbal, quantitative, and analytical, and background information from their first two administrations were considered. Although an attempt was made to be fairly comprehensive in identifying all repeaters fitting the above qualifications, it is certain that not all repeaters were selected. Individuals who changed their names and those who made substantial errors in gridding identification information on their answer sheets were not identified as repeaters.

The sample can be broken down into two groups: the GRE examinees who took the previous version of the General Test at least twice in 1980-81 and have no General Test scores on record earned prior to October 1980 (Old-Old); and the GRE examinees who took the General Test for the first time in 1980-81 and the second time in 1981-82, i.e., the old version first and then the new version (Old-New).

Table 1 indicates which pairs of administrations constitute each of the two samples, which forms were administered, and how many examinees took the pair of forms. The number of months between the administrations are indicated in parentheses.

Insert Table 1 About Here

Method

Using data from the Old-Old group, second scores were predicted from first scores, functions of first scores (e.g., first score squared), and demographic data. Appendix 1 describes the variables that were used. The resulting prediction equations were cross-validated twice. Equations predicting second scores in the Old-Old group were cross-validated in a hold-out sample from the Old-Old group and in the Old-New group. Within each cross-validation the predictors were identical in meaning and weighting. If the cross-validation groups were random samples, any differences between the distributions of residuals is a reflection of a change in scale, a change in the construct(s) underlying the scale, or sampling error. To the extent that all variables that accounted for significant differences between the samples were accounted for in the analyses, essentially, this will remain the case.

RESULTS

Table 2 presents the means, and standard deviations of the residuals and the correlations of predicted and observed second test administration scores for those examinees in two cross-validation groups: the first, a random hold-out sample from the Old-Old group from which the prediction equations were derived, the second, the Old-New group. The Old-Old cross-validation sample serves as a baseline for comparing the residuals in the Old-New cross validation samples. Although we know the expected residual in a randomly selected cross-validation group will be zero, these data allow us to get a handle on how much deviation from the expected zero is reasonable, and also provides us with an estimate of the expected standard deviation of the residuals and the expected multiple correlation.

Insert Table 2 About Here

In predicting the second administration scores, the verbal mean residual in the Old-New group appears slightly low the quantitative mean residual appears moderately low, and the analytical mean residual appears very low. The standard deviations of the residuals for the verbal and quantitative measures appear in line with those from the hold-out cross-validation group, but the standard deviation for the analytical residuals appears considerably larger than that for the hold-out group. The relationships between the standard deviations mirror those between the correlations between predicted and observed scores. The correlation between predicted and observed analytical scores in the old-new group is very low, .79, but this is in keeping with the low internal consistancy estimated reliability of the first several forms of the new analytical measure.

Scrutiny of the residual plots from the predicted score analyses provides evidence regarding possible change in the constructs underlying the scales. Figures 1, 2, and 3 present plots of the mean residuals for the verbal, quantitative, and analytical measures, respectively, for each 50 point interval of predicted score: 200-250, 250-300, etc. No data are presented for any score grouping for which there are fewer than 30 predicted scores. These points are joined by a cubic spline fit to provide an estimate of the mean residuals throughout the predicted score range. Two additional lines surround the line of mean residuals. The upper line is the cubic spline fit connecting the points that are one standard deviation (based on the data within each individual score grouping) above the mean residuals. The lower line is like the upper, only connecting the points one standard deviation below the mean residuals.

If scores on the old-format test and the new-format test were completely interchangeable, one would expect that for each measure the mean residuals would be zero throughout the predicted score range, and the standard deviation of the residuals would be constant. Minor perturbations might be due to sampling error. More serious deviations might be caused by one or more of three factors:

1. one or more important variables were left out of the regression analyses;
2. the original equating of the new-format test to the old-format test was not totally appropriate; or

3. the construct(s) underlying the measures have changed.

It is unlikely that important variables were left out of the regression analyses. Data on more than 100 potentially useful variables were input into the regression. There was little increment in variance explained beyond the fourth or fifth predictor selected, yet 13, 22, and 20 predictors were selected as statistically significant for the verbal, quantitative, and analytical measures respectively. Since GRE annual populations are fairly stable from year to year, and the percentile ranks for comparable verbal and quantitative scores in 1980-81 and 1981-82 matched very closely, it is very unlikely that there could be more than one or two scaled score points of error in the equating of those two measures across the change in test format. Thus, it is likely that most of the variation from the expected pattern of residuals is due to changes in the underlying constructs. Since the content of the verbal and quantitative measures did not change, the change in construct is likely to be related to the change in speededness or the change to rights-scoring instructions. The new analytical measure, on the other hand, was not equated to the old analytical measure, did show substantial changes in percentile ranks of numerically identical scaled scores, and had removed from it two of the four item types that were in the earlier measure.

Figure 1 shows that for the verbal measure, the mean residuals were positive for the lower half of the predicted score range (200-400) and negative for the upper half (450-700). Only at the lowest part of the score range (200-250) was the mean residual more than slightly positive. The figure shows, for example, that examinees for whom scores between 600 and 700 were predicted, on the average, scored 20 to 25 points lower than predicted. Throughout most of the predicted score range, the mean residual was strongly linearly related to the predicted score. The standard deviations of the residuals are very consistent (about 52-56) throughout most of the score scale (300-700). At the lower end of the scale (200-300) they are slightly lower (48-49). All in all, the residual pattern for the predicted verbal scores indicates a small but consistent change in the sources of variability underlying the verbal measure.

Insert Figure 1 About Here

The three major changes in the verbal measure were the shortening of the test from 80 to 76 items, increasing the time limit from 50 to 60 minutes, and changing from formula-scoring instructions to right-scoring instructions. The first two changes were instituted to decrease the speededness of the test. This might result in a pattern of residuals like the one in Figure 1. Test items are, in general, ordered by difficulty, and so typically only the more able examinees (those who might have been able to answer some of the last few difficult items) earn scores that are affected by speededness. On the less speeded new format verbal measure, the more able students would be more likely to have had their scores underpredicted, and thus they would be more likely to have negative residuals.

Alternatively (or in addition), a pattern like the one observed might be the result of a greater propensity toward guessing under rights-scoring instructions for higher scoring examinees. This might be due to lower scoring

examinees, particularly nonnative English speakers, being less able to read and understand the instructions. Alternatively, some lower scoring examinees might be less able to break any existing psychological set against guessing that might have been imposed by previous GRE test-taking experiences under formula-scoring instructions.

Figure 2 presents the residual bands for the quantitative measure, based on the prediction equation for second scores derived in the Old-Old group. Throughout the middle of the predicted score range (400-650) the mean residuals are slightly negative (about 5 to 10 points). At both ends of the scale (250 to 400 and 650 to 750) the mean residuals are moderately negative (10 to 20 points). The standard deviations of the residuals were largest (69-70) between predicted scores of 350 and 450, and decreased toward the high end of the scale (63 from 500 to 600, 56 from 650 to 700, and 49 from 700 to 750).

Insert Figure 2 About Here

Overall, the residual pattern for the quantitative measure shows consistent over prediction of quantitative scores on the new-format General Test. The only substantial change in the quantitative measure was the change to right-scoring instructions. If the self-selected group of repeaters was less likely to guess under the old format formula-scoring instructions than typical examinees, but under rights-scoring directions was just as likely to guess, a residual pattern like the one observed would have occurred. Assuming that many repeaters undergo additional studying or coaching before taking the test a second time, it is possible that they would learn to supress any reticence to guess that they had, particularly under circumstances where they could eliminate one or more of the distractors. This hypothesis does not explain why the standard deviation of the residuals would be lower for the higher predicted scores.

In Figure 3 the residual band for the analytical measure presents a much more extreme picture than do Figures 1 and 2. For predicted scores of 250 to 350, the mean residuals are moderately positive (about 15 to 20 points). From 350 to 450 they are moderately negative. From 450 to 700 the mean residuals are extremely negative. The standard deviations of the residuals are highest (76-79) for predicted scores between 500 and 700. Between 350 and 500 they are somewhat smaller (70-73). At the low end of the scale, 250 to 350, the standard deviations of the residuals are smallest (61-62).

Insert Figure 3 About Here

As no attempt was made to equate the new format analytical measure to the old format measure, it is not surprising that the mean residual is different than zero. Nonetheless, the mean difference of almost 30 points is extreme. This extreme difference, and in particular the unusual shape of the residual plot, appears to be a reflection of the interaction between the factorial structure of the measure and shifts in the GRE population between 1977 and 1981. The current measure appears to depend almost entirely on verbal and quantitative factors (Kingston, 1984). Although verbal and quantitative factors were prominent in the earlier analytical measure, it was factorially more complex (Swinton & Powers, 1980; Rock, Werts, & Grandy, 1982). The 1981

(primarily science majors)

GRE population had a greater proportion of foreign examinees taking the GRE General test at domestic administrations (administrations from which the scaling and equating data come) than did the 1977 population. These foreign examinees tend to do better on the quantitative measure and worse on the verbal measure (and it is through those two measures, weighted equally, that the analytical measure has been scaled) than do examinees who are United States citizens.

CONCLUSIONS

The changes to the GRE General Test that were instituted in 1981 had a small effect on the verbal score scale. The effect on the quantitative scale was somewhat larger. These changes are probably attributable to a shift in the dimensionality of the factors underlying these measures, possibly due to changes in speededness (in the case of the verbal measure), or the change from formula-scoring to right-scoring instructions. The magnitude of these effects should probably not affect the interpretation of scores on the verbal and quantitative measures. The effect on the analytical scale was quite large, and was probably due to an interaction between the factorial structure of the analytical measure and a change in the constitution of the population on which it was scaled. Equal scores on the earlier analytical measure and the current measure are not indicative of the same level of developed ability. The analytical measure is still considered experimental by the GRE Board.

REFERENCES

Angoff, W. Scales, norms, and equivalent scores. In Robert L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Kingston N. Reanalysis of the psychometric characteristics of the revised analytical measure of the GRE General Test. Unpublished paper. Princeton, N.J.: Educational Testing Service, February, 1984.

Rock. D., Werts, C., and Grandy, J. Construct Validity of the GRE Aptitude Test across populations -- an empirical confirmatory study (GREB No. 78-1P). Princeton, N.J.: Educational Testing Service, 1982.

Swinton, S. and Powers, D. A factor analytic study of the restructured GRE Aptitude Test (GREB No. 77-6P) Princeton, N.J.: Educational Testing Service, 1980.

APPENDIX

Table A1

Prediction Equation for the Second Verbal Score[1]

Variable	B Weight	Beta Weight
First Verbal score (V)	9.3×10^{-1}	.87
First V^{-3}	4.9×10^{8}	.10
First V^{3}	-1.4×10^{-7}	-.08
First Analytical score	2.0×10^{-1}	.21
Does not communicate better in English	-1.3×10^{1}	-.04
Birth year - 1931 to 1935	2.5×10^{1}	.02
Birth year - 1941 to 1945	1.5×10^{1}	.02
Years since B.A. - 25 to 45	3.5×10^{1}	.02
Degree objective - Ph.D.	7.0×10^{0}	.03
Neither citizen nor resident alien	-1.9×10^{1}	-.02
Major - Biological science	5.4×10^{0}	.02
Ethnic group - Other	-9.7×10^{0}	-.02
Undergraduate school - Private, not church affiliated	5.4×10^{0}	.02
Constant	-4.1×10^{1}	

Multiple r = .9037

[1] Based on stepwise multiple regression in a group of 4,075 examinees who took the Pre-October 1981 form of the GRE General Test both times.

Table A2

Prediction Equation for the Second Quantitative Score[1]

Variable	B Weight	Beta Weight
First Quantitative score	7.1×10^{-1}	.69
First Analytical score	1.2×10^{-1}	.12
First Verbal score	5.4×10^{-2}	.05
Major - Physical science	1.8×10^{1}	.05
Major - Social science	-1.7×10^{1}	-.05
Major - Humanities	-1.7×10^{1}	-.04
Major - Psychology	-1.5×10^{1}	-.04
Major - Education	-1.6×10^{1}	-.03
Major - Physical education	-2.2×10^{1}	-.02
United States Citizen	-2.0×10^{1}	-.07
Female	-1.5×10^{1}	-.06
Ethnic group - Black	-1.5×10^{1}	-.03
Ethnic group - Oriental	1.3×10^{1}	.02
Ethnic group - Hispanic other than Puerto Rican or Mexican-American	-1.9×10^{1}	-.02
Birth year - 1941 to 1945	-1.8×10^{1}	-.03
Birth year - 1960	1.1×10^{1}	.02
Years since B.A. - one or less	5.0×10^{0}	.02
Years since B.A. - 15 to 19	2.1×10^{1}	.02
Does not communicate better in English	6.6×10^{0}	.02
Undergraduate school - Private, not church affiliated	5.5×10^{0}	.02
Undergraduate grade-point average - A	9.1×10^{0}	.02
Undergraduate grade-point average - A-	5.5×10^{0}	.02
Constant	1.2×10^{2}	

Multiple r = .8657

[1] Based on stepwise multiple regression in a group of 4,075 examinees who took the Pre-October 1981 form of the GRE General Test both times.

Table A3

Prediction Equation for the Second Analytical Score[1]

Variable	B Weight	Beta Weight
First Analytical score (A)	9.1×10^{-1}	.87
First A^3	-4.5×10^{-7}	-.29
First A^{-3}	5.1×10^{8}	.09
First Verbal score	2.8×10^{-1}	.24
First Quantitative score	1.6×10^{-1}	.15
United States citizen	2.2×10^{1}	.07
Attended public graduate school	-7.3×10^{0}	-.02
Ethnic group - Black	-1.9×10^{1}	-.03
Ethnic group - Other	-1.2×10^{1}	-.02
Birth year - 1946 to 1950	-1.1×10^{1}	-.02
Birth year - 1952	-1.2×10^{1}	-.02
Birth year - 1958	7.7×10^{0}	.02
Birth year - 1959	8.9×10^{0}	.03
Birth year - 1960	1.3×10^{1}	.02
Female	5.9×10^{0}	.02
Major - Other	1.9×10^{1}	.02
Major - Biological science	6.0×10^{0}	.02
Undergraduate grade-point average - A	7.2×10^{0}	.02
Undergraduate grade-point average - B-	-6.2×10^{0}	-.02
Undergraduate school - less than 1000 students	-1.1×10^{1}	-.02
Constant	-9.5×10^{1}	

Multiple r = .8813

[1] Based on stepwise multiple regression in a group of 4,075 examinees who took the Pre-October 1981 form of the GRE General Test both times.

Table 1

Definition of Study Sample[1]

Old-Old (n = 5,072)

First Administration		Date and Form of Second Administration			
Date	Form	12/80 A2b	2/81 B1a	4/81 B2a	6/81 C1
10/80	A1a	1,187 (2)	1,048 (4)	413 (6)	233 (8)
12/80	A2b		495 (2)	485 (4)	230 (6)
2/81	B1a			445 (2)	239 (4)
4/81	B2a				297 (2)

Old-New (n = 2,353)

First Administration		Date and Form of Second Administration				
Date	Form	10/81 D1	10/81 D2	10/81 D3	12/81 D1	2/82 D2
2/81	B1a	183 (8)	100 (8)	101 (8)		
4/81	B2a	201 (6)	227 (6)	231 (6)	307 (8)	
6/81	C1	239 (4)	246 (4)	238 (4)	261 (6)	119 (8)

[1]Tabled entries indicate number of examinees who were administered the pair of forms and, in parentheses, the number of months between administrations.

Table 2

Predicting Test Scores Across a Change in Test Format -
Means and Standard Deviations of Raw Residuals
In Two Cross-Validation Groups

		Residuals[1] and Correlations[2,3]								
		V			Q			A		
Format	Sample Size	$\bar{x}$	s	r	$\bar{x}$	s	r	$\bar{x}$	s	r
Old-Old	997	1.2	53.7	.90	-.8	62.4	.86	1.7	63.2	.88
Old-New	2353	-5.1	54.1	.89	-9.4	63.2	.87	-28.9	77.3	.79

[1] Mean and standard deviations of the residuals.

[2] Correlation of predicted score with observed score.

[3] For comparison, the multiple correlations in the Old-Old sample (n = 4,075) in which the prediction equations were calculated were .90, .87, and .88 for Verbal, Quantitative, and Analytical, respectively. Likewise, the multiple correlations in the New-New sample (n = 3,062) in which the predictor equations were calculated were .89, .86, and .81 for verbal, quantitative, and analytical, respectively.

RESIDUAL BANDS OF GRE VERBAL SCORES IN THE OLD-NEW GROUP
(PREDICTION EQUATIONS DERIVED IN THE OLD-OLD GROUP)

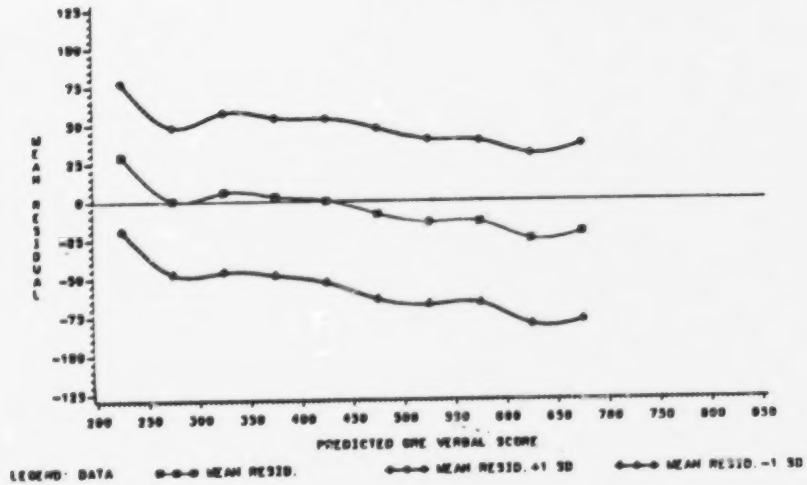

FIGURE 1

RESIDUAL BANDS OF GRE QUANTITATIVE SCORES IN THE OLD-NEW GROUP
(PREDICTION EQUATIONS DERIVED IN THE OLD-OLD GROUP)

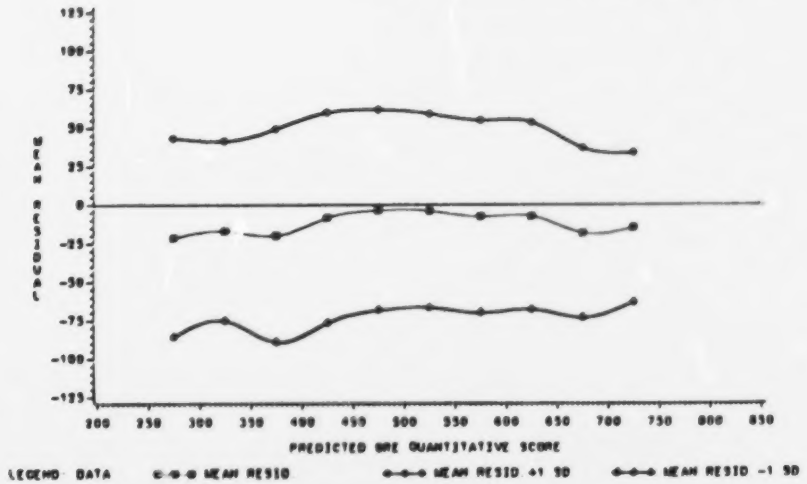

FIGURE 2

RESIDUAL BANDS OF GRE ANALYTICAL SCORES IN THE OLD-NEW GROUP
(PREDICTION EQUATIONS DERIVED IN THE OLD-OLD(1) GROUP)

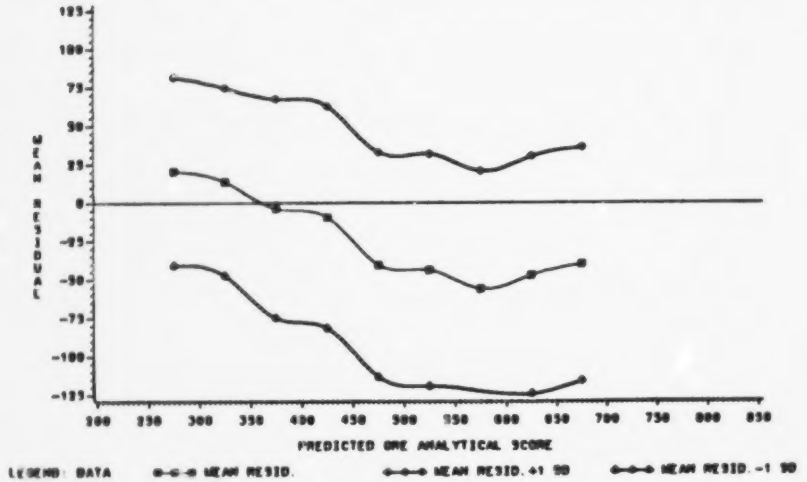

FIGURE 3